自然語言和電腦編程語言的比較

周錫令
北京信息工程學院 教授
xlzhou0421@vip.sina.com

摘要

電腦在處理編程語言方面的巨大成功 和 在自然語言處理方面的舉步維艱 形成了巨大的反差。“比較”是觀察和分析事物的有效方法，把自然語言和人工設計的語言進行一番比較也許能給我們一些工作上的啓示。

作爲資訊傳遞的媒介，目前的電腦語言 和 自然語言 雖然在外表上有很大差異，但是在實質上的確有不少相通或者互相對應的地方。從最初的比較簡單的電腦語言（例如初期的 Basic）到後來越來越複雜的 Fortran, C, C++，我們可以看到把自然語言中的一些機制逐步添加到計算給語言中的迹象。

電腦處理自然語言遇到重大困難，而人卻能應付如裕，是因爲人擁有**知識**（包括社會生活常識以及各種專業知識），並且具有應用這些知識來“解讀”語句的能力。

自然語言是千百年以來爲人際交流的目的而發展起來的。現在，人們越來越多地“上網”，電腦越來越成爲這個語言社會中的一個重要的參與者。人們就會自覺、不自覺地根據電腦的“能力”去總結漢語的規律，反過來影響，甚至改造人們的語言習慣。

一方面，技術向自然語言的深層衝擊，另一方面，自然語言向現代技術靠攏，這兩方面的發展趨勢會和起來，將會是解決“基於語義，基於理解的自然語言處理”這一戰略任務的過程中的一個重要特徵。

自然語言與編程語言之間的相似點

序言

經過幾十年的全球性的努力，以機器翻譯爲代表的電腦自然語言處理工作始終沒有達到人們預想的境界。於是我們竟不住要問：爲什麽電腦處理起“編程語言”來那樣輕鬆自如，可以作好多非常複雜的事情；而在一句普通的自然語言面前卻顯得像一個大笨蛋呢？自然語言和編程語言的本質區別到底在哪里？

自然語言和編程語言顯然有很多地方不同。但是作爲“語言”，兩者都面臨語言使用這所需要的一些要求：

（1）有強大的表達能力（能夠把事情說清楚）
（2）結構化。人的短時記憶容量不多，資訊如果不分層次，無論聽說還是閱讀都會造成困難。
（3）具有簡潔、濃縮表達的機制（使聽說雙方都不覺得囉嗦）。

在這兩種要求的驅動下，兩種語言都會發展出一些機制，這些機制在兩種語言中的表現可能

大不相同，但是會存在某種對應關係。

在文科領域有所謂“比較文學”的行當。考慮到不同民族，不同文化發源地發展出來的文學作品既有各自的特色，又有互通的共性，可以對它們的異同加以比較。事實證明，從這種比較中，可以得到許多有益的啓示。“比較”既然是觀察和分析事物的有效方法，把自然形成的語言和人工設計的語言（電腦編程語言就是應用最爲廣泛的一種人工語言）進行一番比較也許能給我們一些工作上的啓示。

語言的設計者

自然語言是在無數多人群之間的碰撞和交流之間產生、發展、篩選、淘汰之後形成的，好比是“市場經濟”的產物。

電腦**編程語言**（以下簡稱“編程語言”）則是“計劃經濟”的產物。它所使用的辭彙、規則都是事先由一位元“上帝”（語言的設計者）策劃好的。

辭彙及其分類

“詞”是自然語言中的基石，它們是具有語義的最小獨立單位。

在編程語言中對應的東西就是 token。Token 是編譯程序中的術語，它包括外形像英語單詞的 Word，以及“=, +, -, *, /, ==,>, <, (,), ……” 之類的符號。

從資訊處理的角度來看，“詞”和 token 都是“符號（Symbol）”,它們可以被我們“用來”映射到各種實體或者概念上去。根據一個符號所映射到各種實體或者概念的性質來把它們進行分類。

語言學家把自然語言中的詞劃分爲許多類：名詞、動詞、形容詞、副詞、數詞、連接詞、感歎詞、……。我們應該注意到，它們不是在同一級別上的。

名詞 和 動詞： 是最重要的。它們直接反映了我們對世界上形形色色的**事物**以及這些事物之間的**相互作用。**
形容詞、副詞、數詞、則是第二級的。它們只對事物以及這些事物之間的相互作用起**修飾作用。**
剩下來的連接詞、感歎詞、……則是第三級的。它們主要起語法的作用。（用來提示語句內部的結構性資訊，起連接作用、用來表述“詞”與“詞”之間的關係）

與之對應，在編程語言中，可以把 token 劃分爲：

- **命令**：運算符、副程式名。它們對應于自然語言中的**動詞**。動詞的“價”則相當於編程語言中的命令、運算符、副程式所需要的“變元（arguments）”的個數。在編程語言中，有時某些“變元”可以省略不寫，而用預先規定的“缺省值”代替。在自然語言中，也有類似的情況，但是缺省的東西要根據上下文來補充。

自然語言中存在一些“**泛義動詞**”，例如漢語中的“打”，“搞”，“幹”……，英語中的“get”，“take”等。它們的進一步的具體含義要由這些動詞所涉及的物件來確定。例如：“打毛衣”中的“打”應理解爲“編織”，“打籃球”中的“打”應理解爲“玩”，“打開水”中的“打”應理解爲“取得”，等等。

- 在面向物件的編程語言中，也有類似的“動態綁定（Dynamic Binding）”機制：一個函數名或者副程式名字的具體含義要在“運行時”依據所涉及物件在當時的指的類型來決定。
- **資料**：常數、變數名。它們對應于自然語言中的**名詞**。
- **其他**：起連接作用、用來表述 token 之間的關係的符號，例如 **if, then** 等等。它們對應于自然語言中的連接詞、感歎詞等等。

動詞和名詞之間的互換性

在自然語言中，有一個引人矚目的現象，那就是“名詞和動詞之間的呼喚性”。這是因爲，“詞”或者 token 是用來表達“概念”的，而一個概念往往有多種側面。所以自然語言中常常出現用同一個“符號”來表達不同的側面的現象。舉例來說：

“釘”本來是名詞，但是可以轉化，作爲動詞來使用：我把畫**釘**在牆上。

“鎖”本來是名詞，但是可以轉化，作爲動詞來使用：我用鎖把門**鎖**上。

古漢語中這樣的例子更多。“叔”可以轉化爲“認……爲叔叔”，“塵”可以當作“弄髒”，“污染”的意義來使用。

反過來，動詞也往往可以轉化爲名詞使用。比如，“偷”是動詞。但是在“**偷**是不對的。”這句話中的“偷”卻是一個名詞，因爲這句話也可以說成：“**偷這種行爲**是不對的”

在電腦編程語言中，也有類似情況。在 declaration（聲明）中，函數、副程式名表現爲“名詞”，但是在執行語句中，他們就成了動詞。例如在 C 語言中：

```
double x,y;
double sqrt(double);          /* sqrt  在這裏以名詞的面貌出現*/
…….
x = sqrt(y);                  /* sqrt  在這裏以動詞的面貌出現*/
```

詞類的判斷

從語法分析的角度看，對詞或 token 進行分類的依據是它的**語法功能**。從語言使用者的角度看，則是他們的**使用方式**。因此如何對詞或 token 所屬類別進行判斷乃是一件至關重要的事情。

不管在自然語言中，還是在編程語言中，判斷一個詞或者 token 的類別的辦法基本上是兩種：詞典 和 詞本身所攜帶的形態標誌。

詞典中提供的資訊

在編程語言中，爲判斷一個 token 的類別而提供的“詞典”有兩種：

1. 一種是語言中“先驗地”規定好了的**外部詞典**。例如關鍵字和一些保留字。
2. 另外一種是編程人員（用戶）臨時定義的**內部詞典**，這就是程式中的 declaration.

在自然語言中，基本上只使用外部詞典。（在某些文件 [特別是有關技術標準的文件] 中，有時也在文件的開頭部分定義一個“術語”集合，但是很少涉及詞性的問題）。

例如中，市面上提供的**英語**詞典對其中所收羅的詞的詞性都給出了說明。與編程語言不同的地方是：有相當一部分詞具有多種詞性。例如：字串“increase” 既可以當作名詞來使用，又可以當作動詞來使用。

奇怪的是，**漢語**詞典基本上都沒有給出詞性方面的資訊。其主要原因大概是由於漢語中大部分的詞都允許以多種方式使用，也就是說具有不止一種詞性。這種現象在古漢語中標顯得尤爲明顯。漢語中的這種“傳統”的來源可能如下：

漢語是象形文字，最方便給有形的物件起名字。所以至少在漢語中“名詞”最優先地得到發展。在古漢語中，由於爲抽象的動作設計象形字比較困難，所以往往就把名詞直截了當地轉化爲動詞使用：“老”指老年人，老年人應予以“尊重”，所以就把它當作現代漢語中的“尊重”這一動詞來使用。“幼”指嬰幼兒，嬰幼兒需要“愛護”，所以就把它當作現代漢語中的“愛護”這一動詞來使用。於是就出現了“**老吾老以及人之老，幼吾幼以及人之幼**”以及與之類似的“**君君臣臣父父子子**”這類令現代人費解的句子。

詞類劃分的形式標誌

如果能夠從詞或 token 的外形（形式上的）特徵就能判斷出他是屬於哪一類，那麼無論從“書寫者撰寫”的角度，還是從“閱讀者理解”的角度，都能夠大大減少出錯的機會。

在某些編程語言（例如 Visual Basic）中，如果一個變數沒有在任何地方加以聲明,也可以從變數名字的外形上看出它的類型。例如：名字以%結尾的變數是‘整數’， 名字以&結尾的變數是‘字串’，名字以&結尾的變數是‘浮點數’等等。

英語中，在某種程度上也有類似的機制，例如：以 tion, ing 結尾的基本上是名詞。以-lize 結尾的基本上是名詞。以-ful 結尾的基本上是形容詞。以-ly 結尾的大概是副詞。

漢語使用方塊字，沒有辦法添加尾綴，所以沒有這樣的形態標誌。因此大家認爲，這一現象給漢語的電腦處理增加了困難。不過話不能說得太絕對。在某些情況下，漢語還是有“形態標誌”的。例如，在名詞的前面加“很”“還”之類一般用來修飾形容詞的副詞，就是在“形態”上指出：後面的這個名詞已經轉化爲“形容詞”了。例子如：“同學們說我穿這條裙子很**青春**。”，“我們排演的這套節目還是很**生活**的。”，“他比林彪還**林彪**。”。

作用域

編程語言的名字都有“作用域（scope）”問題。這一點在語言設計中都作了明確、毫不含糊的規定。程式被劃分爲模組、分程式，它們是作用域的天然邊界。

自然語言中當然有類似要求。章節、段落的劃分，引號（“ ”，‘ ’）的使用也是名字的作用域的天然邊界。但是與編程語言相比，並沒有硬性的規定，讀者往往要利用生活常識依據“語義合理性”來進行判斷。

指標

自然語言中的指代詞（你、我、他、它等等）好比編程語言中的“指標(point)”。但是自然語言中從不明顯地交待：從現在起，“他”表示“張三”，直到遇見新的聲明爲止。每一個具體的代詞指向何方要根據句子域句子之間的前後語義來聯繫來判斷。

在爲了處理自然語言而爲電腦編制詞典的時候，一個十分重要的問題是：我們把自然語言中的“詞”看成是一個“概念（concept）”還是只看成一個“符號（symbol）”。字面上的一個詞可以對應多個概念，例如“編輯”既可以指一種工作、職業，又可以指以這種工作爲職業的“人”。目前這方面似乎仍存在不同的看法，但是從電腦處理的角度看，當然是看成一個“符號”爲宜。這樣就出現了如何判斷某一個“詞（符號）”指向何種概念的問題。自然語言中使用最多、最具有生命力的詞大多具有多個“義項”。從這個觀點看，**多義項詞** 更接近於編程語言中的**指標**。

語句類型

動作語句

自然語言中，像：

税收人員 向 大家 **宣傳** 税收政策。”

我 **吃**了 一塊蛋糕。”

都表示某一主體採取了某種動作，因而改變了世界。這種句子圍繞著中心動詞“宣傳”，“吃”而展開。這種“動作語句”顯然對應於編程語言中的“命令語句”（賦值語句，副程式調用，帶有副作用的函數調用等。）

如果使用編程語言中的形式來書寫上述兩個句子，結果就是：

宣傳（税收人員，大家，税收政策）

吃（我，一塊蛋糕）

副程式“**宣傳**”有三個變元，所以“**宣傳**”是三價動詞。副程式“**吃**”有兩個變元，所以“**吃**”是二價動詞。

描述語句

自然語言中，描述句以靜態的方式描寫周圍的世界，某種事物的存在，或者它的屬性。在漢語中，最常見的是以“是”爲中心謂詞的描寫句，例如：

她 是 近視眼。（健康）
他 是 小孩。（年齡）
他 是 高個子。（身材）
他 是 工程師。（職業）
……

可見，中心謂詞“是”把句子分成左右兩個部分。右邊部分敍述了左邊部分的某種屬性。至於具體是什麽屬性，完全要依靠讀者的知識來判斷。也許把“是”稱爲“系動詞”就是因爲它的功能只不過是把左右兩部分聯繫起來，或者說，只是指出左右兩部分有聯繫，至於什麽樣的聯繫，則語焉不詳。

由於系動詞只起“語法上分割、語義上聯繫”的作用，因此它往往可以被省略（這時可以認爲中心謂詞是“Φ” ）。

這人 Φ 黃頭髮。
你 Φ 傻冒。

在英語中也有類似情況，特別是在要求句子簡短有力的場合：

You baby!
You silly boy!

編程語言中，與描述句對應的東西是 declaration 以及 declaration 中的“初值語句”，例如：

```
int   x = 3;
John.Hair.Color = Yellow;
```

疑問語句

“你進來好嗎？”，“3 加 5 等於幾？”這類疑問句似乎在編程語言中找不到對應物。其實是有的，就是包括函數在內的“**運算式**”。

電腦程式在運行過程中遇到包括函數在內的“**運算式**”時，就要計算這個運算式的值，也就是向電腦硬體、程式庫、作業系統詢問計算結果。

語境的動態變化

目前我們所使用的電腦都是基於馮·諾意曼模型的機器。其中最具有特點的就是“賦值語句對環境所施加的改變”和隨時用來記錄不斷改變著的環境的存儲系統（寄存器、堆疊、記憶體、外存）。正是由於這一原因，多年以來所發展出來的、針對靜態環境的數學證明方法在“程式正確性證明”的問題上失去了效力。

自然語言中的“動作語句”既然和編程語言中的“賦值語句”相當，它也必然會產生同樣的難題。以當前水平的機器翻譯軟體爲代表的自然語言處理軟體都是“沒有記憶”的，並不把

所處理的語句對環境的影響記錄下來。

如果自然語言處理軟體有這種記憶機制，那麼它在處理以下句子

1. 孫武 是 春秋時代的 軍事家。
2. 他說：“……”。

的時候，就會在處理完第一句後，在“**情景堆疊**”中記住：

現在有了一個最新被提到的人物：春秋時代的 軍事家 - 孫武

於是在處理第二句時，就知道句子中的“他”就是這個“孫武”，由於他是春秋時代的人，所以“說”要使用“過去式”。

當然，這個例子太簡單了。“**情景堆疊**”應該怎樣設計，尚有待於探討。但是這個問題是電腦理解自然語言時避免不了的。

小結

從以上討論可以看出兩點：

1） 作爲資訊傳遞的媒介，目前的電腦語言 和 自然語言 雖然在外表上有很大差異，但是在實質上的確有不少相通或者互相對應的地方。

2） 從最初的比較簡單的電腦語言（例如初期的 Basic）到後來越來越複雜的 Fortran, C, C++，我們可以看到把自然語言中的一些機制逐步添加到計算給語言中的迹象。

自然語言與編程語言之間的區別

這個問題之所以值得研究是因爲這一研究有可能幫助我們搞清楚電腦在處理自然語言時所遇到的困難的本質 以及克服這些困難的途徑。

語句結構分析時語義的介入必要性

以上我們討論了自然語言與編程語言之間的相似之處，也順便指出了兩者之間的一些區別。但是上面所說的這些區別不一定都是本質性的、或者根本性的。兩者之間本質性的區別表現在：自然語言處理在許多階段的工作中都**需要涉及語義，需要語言之外的生活常識、社會與自然科學知識**的支援才能完成。以下是其中主要的幾個方面。

漢語句子的分詞和詞性標注

分詞雖然可以在大部分情況下利用辭典和匹配的方法得到正確的答案，但是有時也需要運用語言之外、語境方面的知識。“乒乓球拍賣完了”應該讀成“乒乓球 拍賣 完了”還是“乒乓球拍 賣 完了”的問題就是一個典型的例子。

編程語言中的 token，除了“指標”以外，基本上都只屬於一種類別，沒有任何含混之處。然而，自然語言中的詞卻可以有多種詞性，漢語尤其如此。但是在一個具體的句子中，每一個詞的詞性確是唯一的，因此就存在著如何根據這個詞的周圍的環境來判斷其詞性，這就是**詞性標注**問題。電腦可以從統計規律出發“胡蒙亂猜”，並且期望能夠在大多數情況下得到正

確的結果，但是像上面所舉例子：“把收集到的資料記錄在資料庫**記錄**中”就很難保證正確判斷其中的兩個“記錄”的詞性。

句子結構分析

我們都知道，句子從外表看只是一維的“字詞流”，但它是有它的內部結構的。分析或者理解一個句子的時候，第一件事就是要從它的一維的字詞流中提取句子的內部結構。正是在作這件事的時候，出現了自然語言和被稱之爲“形式語言”的編程語言的本質區別。

在對電腦編程語言進行語法分析的時候，由於語句中每一 token 均由確切的含義，語言的語法大多限定在上下文無關文法的範圍之內，編譯器只需要擁有語句中的“形式標記資訊”和在語言內部預先定好的一些規則就可以完成任務，而且所得結果是唯一的，用不著利用與**外部世界**有關的知識。編譯器在語法分析結束以後進入代碼生成階段時，才要涉及語義問題。

反之，在分析自然語言中的句子的時候，即便是分詞、詞性判斷等標誌性資訊均已完備的情況下，也往往要借助“語義合理性”方面的判斷。而“合理與否”則要依靠語言之外的生活常識、社會與自然科學知識。舉例來說，

她 想　穿 將軍 的 大衣。

她 想念 穿 大衣 的 將軍。

著兩句話的形式都是：**“名 動 動 名　‘的’　名”**的格式。但是內部結構完全不同。我們在閱讀這兩句話的時候，可以毫無困難地分別把它們理解爲：

她 想穿 （將軍 的）大衣。

她 想念 （穿 大衣 的）將軍。

這是因爲我們有“人可以穿大衣，但是大衣卻不能‘穿’人”的常識。沒有生活常識以及運用這種常識的能力的電腦，僅僅憑語言知識和語句中的形式標誌顯然是沒有辦法完成這種句子的結構分析的任務的。

上下文相關

自然語言中即使其最簡單的“義項選擇”問題，也往往要根據上下文、乃至整篇文章以外的知識（社會生活常識、自然科學與專業知識）來判斷。

例 1：漢語中，“**躊躇**”是“猶豫不決”的意思（如說“躊躇不前”）。但是如果後面緊跟著“志滿”（躊躇志滿），意思就來了一個一百八十度的大轉彎，變成“從容自信”了。

例 2：如上所述，漢語中的“**打**”字是一個泛義動詞，在“打毛衣”，“打籃球”，“打水”三種說法中分別理解爲“編織”，“玩”，“取得”。這一要求可以通過對“詞語搭配關係”的考察來解決。

在電腦語言中，也有類似情況。例如用來打開文件的 open 就是一個泛義動詞，在打開某一具體文件時，到底要使用哪一個 application，則要看這個文件的類型是什麼？例如，如果檔

案名的尾碼是 .doc， 就使用 MS Winword, 如果檔案名的尾碼是 .txt， 就使用 Notepad.

然而， 在自然語言中，有時更複雜一點。例如：在以下這句話中：

漁民用竹篙**打**水，想把魚趕進漁網中去。

“打”自由應理解爲“**擊打**”的意思。而在

漁民用木桶**打**水洗菜做飯。

中的“打”又依然應理解爲“**取得**”的意思。換言之，一個泛意動詞的語義不僅取決於它的賓語，有時還得看看“施主”和“工具”是什麽？

葉聖陶先生在他所寫的一篇文章《據理論而言》中也提到一個例子：

> 解放前“某大學招考新生，國文的作文題目是《防民之口甚於防川論》。很有些應試的同學解錯了題目，做出來的文字牛頭不對馬嘴。”

之所以會這樣，是因爲“防民之口甚於防川”這句話可以有很多解釋：

防民之口**其困難**甚於防川
防民之口**其重要性**甚於防川
防民之口**其……**甚於防川
防民之口**其危險**甚於防川

這些“應試的同學”大概都理解爲前面兩種。如果他們看到《古文觀止》裏收錄的《國語》裏的那篇“召公諫曆王止謗”的文章中這兩句話的**下文**：

防民之口，甚於防川。**川壅而潰，傷人必多，民亦如之。**

就知道應該按照相反的方向來理解了。

自然語言中這種“必須依靠文章以外的**知識**來解讀 作爲符號的“詞”的指向”的要求，給電腦處理自然語言的工作帶來了重大困難。

自然語言的“簡略性要求”

請觀察以下這幅漫畫：

雖然這只是寥寥幾筆，離開相片或者真實的人像相去甚遠，但是當今社會上大多數中國人都

能一眼看出這就是 聶衛平。因爲它抓住了 聶衛平 有別於其他人的特徵。不過，能夠做到這一點的人一定是預先看見過 聶衛平 的相片、錄影或者他本人。換句話說，看得出這幅漫畫畫的是誰的這些人，必須在腦子裏預先已經有了與聶衛平面貌有關的“**知識**”，這幅漫畫不過是“勾起了”他們腦海中已有的印象而已。

類似地，人們講話時的首要目標並不是要去描述和傳遞巨細無遺的情景，而只是在對方已有的知識背景之上做一些補充，使得對方把你的詞句和它已有的知識加起來，能夠湊成一幅完整的圖景。超過這一要求時，別人就會覺得你囉嗦；過於簡單則會使對方誤會或者不知所云。從這一點看，自然語言理解過程得著一側面和“猜謎語”是相通的。漢語歷來有講究“凝練”的傳統，因此我們看到的漢語句子常常只是“海面上的冰山”。

最普遍的例子是“**名詞的串接**”。例如：“**傻瓜相機**”這個片語可以有多種理解，看你省略的字眼是什麼：“傻瓜發明的相機”，“傻瓜才使用的相機”，“爲傻瓜而設計的相機”……。“**魯迅回憶錄**” 這個片語也可以有多種理解：“魯迅自己撰寫的回憶錄”，“某某人撰寫的關於魯迅的回憶錄”。“**海鮮醬油**” 可以理解爲：“以海鮮爲原料製成的醬油”，也可以理解爲：“吃海鮮時使用的醬油”。讀（聽）者需要根據當時當地的語境和自己的生活經驗和常識來選擇其中的一種。

標語、標題中的字眼也是常見的“簡略說話法”的例子。比如說：我們在計程車的座位上常常看見的：

停車等候請付押金

實際上是

停車等候　請　　付押金

如果乘客中途下車辦事，要求司機　　，　乘客預先向司機支　　。

顯然，只有把被省略的詞語適當地補上之後，才談得到句子結構的分析。

可是，除了爲了避免囉嗦、使語言簡潔的目的以外，還有一種情況是“**有意模糊**”。我們注意到，實際生活中，我們使用語言的方式可以分成兩大類。

- 一種是力求精確、完整地傳遞資訊。科學、法律方面的文獻可以作爲這一類的代表。
- 另一類是：有意不把話全部說明白，給讀者留下想象的空間。

編程語言當然屬於第一類，而文藝領域的文章大多屬於第二類。文藝作品的作者在撰寫他的文字時，一方面希望利用這些文字傳遞資訊，但與此同時，還期望勾起讀者的聯想，希望讀者利用自身的知識、處境、心情……進行補充。因此這種文章就要寫得“含蓄、凝練”，給讀者留有“**再創造**”的空間。讀者的主動參與往往調動了他的濃厚興趣。（這也是爲什麼比較

含蓄的音樂作品、廣播劇可以反復聆聽；而一覽無餘電視劇很少有人看兩遍的原因。）

語句分析的“終極”在哪里？

電腦語言的編譯器在分析電腦程式中的語句時，最終總是要分析到語法規則中所規定的“終極符（terminator）”爲止。

在自然語言中的情況比較複雜。漢語句子中的終極符顯然不是漢字，但也不一定是“詞”。在“他胸有成竹”這句含有成語的話中，“胸有成竹”應該是終極符，因爲你不能再繼續把它進一步分解了。再分解就真的變成“他的胸腔中有一根竹子”了。同樣地，“大家奮力力救火”中的“救火”也是終極符，因爲你一定不能再進一步把“救火”當作“動賓短語”去進一步分析。“救火”中的“救”與“火”不過是由“搶**救**生命與財産於烈**火**之中”這句話裏提取出來的兩個“特徵字眼”而已。

英語沒有漢語中的分詞的麻煩，但是也有類似問題。許多介詞短語是一個整體，不能繼續分解。比如說，當討論某人的家庭情況時，有人回答：“He is on the street.” 這“on the street”就應該整體地作爲表示“無家可歸”的一個片語來理解。

話又說回來，要是孩子的父親問母親：“孩子上哪兒去玩去了？”母親若回答“He is on the street.”那就表示這孩子真的是在街上玩。可見，一個片語應不應該作爲“終極符”，還得要看語境。這進一步增加了自然語言理解的複雜性。

規則與習慣

電腦編程語言中，只要是符合語法規則的句子就都是合法、並且可以使用的。但是在自然語言中，卻要看‘習慣’。你雖然在離開家時可以說：“我用‘鎖’把門‘鎖’上”，但在回家時，卻不能用類推的方式說：“我從兜裏掏出‘鑰匙’來‘鑰匙’門”。因爲（社會上）沒有這種說法！（這大約是因爲後面這種說法太‘繞口’）。反之，看起來雖然不大合乎語法，但是大家說慣了，卻是允許的。

隱含與聯想

我們可以把“自然語言”進一步劃分爲“**技術語言**”和“**文化語言**”兩種。前者的例子如：技術資料，法律文件等。後者的例子如：詩詞、小說等。兩者的主要區別是，“文化語言”常常含有“言外之意”，並且引起或者要求讀者（聽者）發揮“聯想”的主觀能動性。這個要求顯然離開當前電腦的水平更加遙遠，所以在此不再展開討論。

繞口與否 及 聽覺特性

自然語言首先是爲了“說和聽”，其次才是爲了“寫和讀”而發展起來的。這樣一來，句子的順不順口，文章的“抑揚頓挫，音調鏗鏘，朗朗上口”就成了一種不可忽視的要求。有時候，爲了滿足這一要求，甚至會“以聲害意”，以顛倒語序爲代價。這種現象在詩歌中當然比較多（例如杜甫的“名豈文章著，官應老病休”），但是在口語中也有出現。魯川教授就曾經指出：我們從來不說“**制 電影片 廠**”而是說成較爲順口的“**電影 製片廠**”，是因爲“2+3”比“1+3+1”結構讀起來更爲順口,雖然前者在語法上較爲合理。由於同一原因，漢語中偏好使用長度爲四個字的短語。（所以漢語中的成語絕大多數是四個字的，如果不夠四個字，往往也要加進一兩個虛字，湊成四個字。）可是，過去我們經常說的“**百分之百**”在說廣東話的人群中遇到了困難，因爲廣東話中沒有‘zhi’這個音。於是他們把“百分之百”改說成“**百分百**”。又由於改革開放以來，以廣東省爲代表的中國南方地區在經濟發展上處於

有影響的位置，使得電視臺上的主持人和播音員也都改說“百分百”了。

反過來，電腦編程語言完全是爲了“寫下來供電腦執行或者供人閱讀”的。所以不存在上述需要朗讀而帶來的“聽覺性”問題。

視覺特性

漢語使用象形字，因而在“字”和“詞”這一級就有視覺特性的問題。我們初中時就讀過李煜的詞《浪淘沙令》中的“簾外雨潺潺，春意闌珊。”後來看見老舍先生批評這種句子說：從來就不知道雨怎麼個“闌珊”法，春意又怎麼個“闌珊”法，覺得老舍先生講得對。可是又覺得，雖然從來沒有去詞典上查過“潺潺”是什麼意思，但是當初讀李煜的詞時並沒有覺得費解。但講不出道理。

最近看到王蒙在《道是詞典還小說》這篇文章中解釋了這個問題：

> 問題不在於“潺潺”本身的含義，對於我來說，“潺潺”的說服力在於字形中放在一堆的六個“子”字，它們立即使我想起了流水上的絲皺般的波紋。從上小學，我一讀到“潺潺”二字就恍如看到了水波。

除了“字”和“詞”這一級以外，自然語言在段落、章節的劃分上也有“語義”上的含義，並且隱含著各種代詞的“作用域（Scope）”劃分。

電腦編程語言中，雖然在“段落”一級上應該力圖在視覺上顯示程式的嵌套結構。以利於“人”的閱讀。但是這種做法對電腦好毫無意義。“作用域（Scope）”劃分也應該有明確的 declaration 來規定，而不能是隱含的。

小結

傳統的電腦語言 和 自然語言的上述差別是造成電腦處理自然語言遇到重大困難的主要原因。人之所以對自然語言應付如裕，是因爲人擁有**知識**（包括社會生活常識以及各種專業知識），並且具有應用這些知識來“解讀語句（猜測語義）”的能力。

事實上，即便是人，由於文化程度和專業領域的不同，如果他聽（讀）到語句在解讀時所需要的知識超過了他的知識範圍，也會產生理解上的困難。因此，人們在通過語言相互交流時，還需要遵從**某種協定**。

語言交流中的普適性協定

人與人之間通過語言進行交流時，人人都自覺或不自覺地遵守以下普遍使用的協定：

1. 發話一方首先要判斷對方的語言能力，並據此選擇可以使用的詞語和句式。如果對方是小孩，使用的詞語和句式就和對方是成人時不一樣。
2. 其次要估計對方的知識水平、專業範圍。不能使用對方無法理解的詞語或概念。例如，如果對方是考古專家，你就不能和他深入地探討其他專業領域的問題。
3. 發話方所進行的以上的兩種估計不可能完全正確。因此，聽（讀）方要隨時判斷對方所說的句子是否在自己所能領悟的範圍內。如果不是，就要要求對方解釋或者換一種易懂

的說法。

4. 發話方所說的話如果小有紕漏，聽話方應該有能力糾正，“正確地加以理解”。

在人和電腦講話時，出現了新的問題。一般來說，人（用戶）沒有和電腦通過自然語言講話的經驗，因此不知道怎樣估計電腦的語言能力和知識水平。在這種情況下，就要研究**如何讓電腦主動向對方以適當的形式介紹自己的接受能力**。這是一項具有挑戰性質的任務。

上述協定中的第三點格外重要。傳統的“全自動”機器翻譯程式會忽略了這一要求，可以說是一個誤區。

兩栖語言-面向電腦的“受限（標注）漢語”

要求電腦擁有一般人所具有的**知識**以及 具有應用這些知識來“解讀”自然語言的能力，目前還不現實。因爲這種要求有點像希望對古代社會生活情況一無所知而又僅有小學語文水平的人去閱讀處處引用“典故”的古籍那樣。不過，語文老師早已想到了解決這個問題的方法：**改寫** 和 **添加注解**。

這一思路當然也可以用來幫助“沒有社會生活常識和語言水平很低的”電腦。這就是提出“面向電腦的受限（標注）漢語”這一概念的由來。

問題在於：誰來進行這種 **改寫** 和 **添加注解** 的工作？答案是：開發“受限（標注）漢語寫作器”，讓電腦幫助“人”來作這件事。

北京資訊工程學院在承擔國家“九五”重點科技攻關專案“**中文資訊處理及產品開發**”中的“**受限漢語技術及其產品開發**”專題工作所開發的“**受限漢語寫作器**”最後產生出來的語言文本有兩種表現形式：

- 給人看的文本。這種文本從外表上看和一般的文章沒有什麽差異。
- 給電腦看的文本。這是在上一種文本中添加了供電腦閱讀的“標誌”

這種帶“標誌”的文本 和 當前正在如火如荼地發展著的 XML 語言 有密切的聯繫。這些“標誌”之除了每個語句中各個語言成分的“語義角色”和它們在“語法意義上所處的位置”。有關情況請參見有關資料，再此不予贅述。

可是，不管怎麽說，**改寫** 和 **添加注解** 終究是額外的工作，而“用電腦處理自然語言”的初衷本來就是爲了節省人力。這兩者豈不是“南轅北轍”嗎？爲此，我們應該從**用戶需求**的角度來探討一下。

兩種用戶需求

資訊收集者

在互聯網、資訊爆炸時代，如何利用電腦從浩如煙海的以自然語言表達的文章中找出自己所關心的、對自己有價值的文章，是很多用戶面臨的問題。

由於電腦在自然語言理解方面尚未取得有價值的進展。因此用戶只能在資訊判斷的準確性、詳盡程度方面降低要求。

資訊發佈者

反過來說，資訊發佈者當然他所發佈的資訊能夠以最大的精確性得到盡可能多的人群（包括不同國別的人群）的注意和理解。爲此，他可以而且願意付出一些代價，以“受限”和“標注”的方式發佈自己的資訊。

制定“受限（標注）漢語”規範的思路

我們這裏所說的“受限（標注）漢語”是“面向電腦的受限（標注）漢語”。換言之，付出“簡易”的代價之後，期望達到的目的是**使“智慧相當低的”當代電腦有能力加以處理；與此同時，在人類讀者面前依然保持原來自然語言的風貌**。

顯然，“面向電腦的受限（標注）漢語”是否有實用價值，取決於：

- 我們能否爲這種“受限（標注）漢語”設計出既有充分表達能力，又便於電腦處理的語言規範。
- 開發出一種足夠好的幫助人們書寫這種語言的工具---

大家知道，法律有兩種表述形式：

- **案例法**（Case Law），屬於“大陸法系”，“羅馬法系”
- **法典法**，屬於“英美法系”或“普通法系”。

自然語言的規範也有這樣的兩種。小孩學習自己國家的母語時肯定是根據“案例法”。中國人過去學習英語的傳統辦法是“法典法”。自從《英語 900 句》一書風行以後，很多人體會到了使用“案例法”學習外語的好處。

剛開始學習編寫電腦程式的時候，我們多半是對照著程式語言手冊上的規定一句一句地寫（法典法）。對於年齡大、記憶力差的人，常常感到許多細節記不住，就希望找一個或者一段別人已經寫好的程式拿來“修改”，覺得既省事，又少出錯。（案例法）

那麼我們制定受限（標注）漢語規範的時候，應該採取哪一種辦法呢？

我們曾經想過，能不能定義一個漢語子集，使得這個子集內的語句像電腦編程語言那樣滿足某種上下文無關文法，這樣一來，電腦處理編程語言的技術不是就可以用來處理這種受限（標注）漢語了嗎？。（這就是一種“法典法”的道路。）

這條道路給我們帶來的困難是：自然語言的內涵比編程語言要複雜得多，爲它設計一個滿足上下文無關文法的子集有“老虎吃天，無從下嘴”的感覺。更爲重要的是，即使設計出了這樣一套語法，一般人也一定無法使用。理由很簡單，因爲人們學習自己的母語時，不是走的這一條道路。覺得多數說漢語的人根本沒有“漢語語法”的概念。

因此合理的辦法是採用“案例法”。也就是在電腦中存儲足夠多的模板和例句然後引導用戶“依葫蘆畫瓢”。這種辦法用戶容易接受，更爲重要的是：由於模板是事先定義好的，它的內部結構，各部分之間的語義以及相互之間的聯繫也是預知的。因此用戶在“選句填詞”的過程中，無需用戶費心，系統就可以自動產生大量的標誌性資訊。

但模板法也有困難。首先是如何進行引導？其次，自然語言（特別是漢語）語句中的一部分又可以以遞迴的方式展開爲另外一個句子。對於“一級”語句，我們可以用模板引導的辦法加以限制，但對於以遞迴的方式衍生出來的第二、第三級句子就沒有辦法再這樣做。限於篇幅，這兩個問題留在別的地方去討論。

漢語規範 與 漢語.DTD

我們經常在報刊上看到希望使用“規範漢語”的呼籲。但是，怎樣的漢語才算是規範，卻沒有一個客觀的標準，於是產生了“你說我的語言不規範，我卻認爲很規範”的爭論。考慮到語言的進化過程是不可阻擋的，而在進化過程中又同時進行著篩選、淘汰的過程，這個問題的解決恐怕還得依賴於技術發展的反作用、“技術標準”以及相應的標準的定期更新。

舉例來說，長期以來，在我國人名、地名用字沒有標準，沒有範圍，有些地區甚至可以自己造字。但是在資訊處理用漢字國家標準 GB2312 頒佈以後，許多家長再給自己的孩子取名的時候，都知道最好在 GB2312 所收羅的漢字集中選字，否則的話，在電腦中找不到孩子的名字將使孩子終生受累。

類似地，電腦中的“漢字形檔”，OCR 軟體，中文拼音輸入方法，教育部在小學推廣的“筆劃”輸入法，天天在對廣大的使用電腦的人群實施著規範化的培訓。

由此推想，如果有了電腦能夠接收和處理的漢語規範，並且得到實施，那將對鼓勵人們使用規範化的語言起很大的推動作用。因爲，如果你不遵從者以規範，不僅有一種客觀的手段來判斷你的“錯誤”，而且你寫的文章電腦無法處理，因而你的思想就不容易通過網路得到迅速傳播，也不易得到他人的廣泛理解。

那麼，我們所設想的這種“規範”的形式應該是怎樣的呢？也許第一步最合適的形式是 DTD 文件。

前面說到，北京資訊工程學院所開發的“**受限漢語寫作器**”最後產生出來的給電腦看的文本中添加了供電腦閱讀的“標誌”。這種帶“標誌”的文本 和 當前正在如火如荼地發展著的 XML 語言 有密切的聯繫。這些“標誌”之除了每個語句中各個語言成分的“語義角色”和它們在“語法意義上所處的位置”。但是這套“標誌“只是負責這一專題的開發小組的看法。其合理性尚有待於進一步改進和驗證。

一個十分重要的問題是，應該爲這種“標誌”建立一套大家公認的標準，並且把它體現在.DTD 文件中。我們知道，在數學領域已經形成了針對數學領域的標誌語言的 MathML.DTD。對於漢語，也應該形成類似的 DTD 文件。我們知道，有關部門曾經不止一次地組織力量編撰“漢語語法”，但是遇到的困難不少。從電腦處理的角度出發，首先在“標誌”的種類與定義上達到共識，形成一個大家公認的“漢語.DTD”文件，將有助於電腦對漢語的處理 和 漢語規範化工作。

對寫作工具的要求

交互過程的必要性

由於目前還不可能把一般人頭腦中擁有的全部知識以適當的形式裝入電腦，並使之具有運用這些知識把語句中所缺省的資訊補充進去的能力，因此必須通過人機對話模式，在寫作過程中向電腦“交代、補充”有關資訊。

減少用戶干預的工作量

上述這種交互過程顯然是令用戶厭煩的。因此技術工作的目標就是要設法使這種額外的工作量減到最小。

- 雖然目前還不可能把一般人頭腦中擁有的全部知識以適當的形式裝入電腦，但是有些比較簡單的知識還是可以以適當的形式裝入機器的。（例如：在電腦中保存詞語搭配規則、捆綁規則、增大機內詞典所收容資訊的粒度等等。）
- 在“詞性”，“義項”的選擇方面，如果不能進行有理智的分析和推理，就採用基於統計的猜測方法。只要大部分情況下猜測得正確，就可以大大減少要求用戶干預的次數。

討論

進行上面進行的“比較”當然是爲了幫助我門弄清楚今後的努力方向。

技術邊界

一個“個人”的知識總是有限的，它所能處理的語言資訊在內容、形式上都會有限制。因此，沒有任何一個人敢說他能看得懂任何一篇文章。人尚如此，“機”何以堪？對於一台資源有限的電腦，當然也是這樣。回想起在中學裏學“幾何”課的時候，老師就曾告訴我們：僅用圓規和直尺，是不可能在理論上找出三等分任意一個角度的辦法的，因此不要白費勁去做那些不可能做到的事情。這事啓發我們，在電腦中文資訊處理的領域裏，也不能一味“埋頭拉車”，應該先問一問，使用電腦處理自然語言的技術邊界在哪里？

根據目前中文資訊處理的狀況看來，我們無法提出一種決然分明的技術邊界，如果畫三個由大到小的圈子：

1）原本的、不斷地進化著的 **漢語**
2）從 **漢語** 中剝離掉 **文化語言**，只研究 **技術語言**（也就是沒有‘言外之意’，不使用隱喻、聯想和暗示的語言。
3）對**技術語言**在所用的辭彙、句型方面進一步加以限制而得到的 **受限漢語**。

顯然，畫的圈子越小，電腦所能達到的理解的深度會越高。另一方面，也應該考慮到，不同性質的處理對“理解”提出的要求是不同的。

不同性質的處理對“理解”提出的不同要求

“自然語言”和人類積累的知識**相共生，同發展**的。只要電腦擁有的知識和運用知識的能力沒有達到“人”的水平，我們便不可能要求電腦在總的方面達到和人一樣的處理自然語

言的能力。

但是，不同性質的處理對“理解”提出的要求是不同的。

機器翻譯

機器翻譯工作的一個特點是：有些時候，翻譯者即使並不懂得（或者不完全懂得）句子的內容也照樣可以“按字面”進行翻譯，而聽者也能正確理解對方在說什麼。這種情況常見於外語專業的畢業生爲兩位同行的專業（例如電腦軟體）工作者進行技術翻譯的時候。顯然，電腦也可以在類似情況下用“字串代換+簡單的重新排序”的辦法達到“正確”翻譯的目的。

不過這種“有利情況”並不總是發生，更多的（大約 70%）情況需要翻譯者對談話的內容有深入的理解。絕大多數“四個字的中國成語”就不能按字面翻譯成外語。這大約就是全自動翻譯系統的正確率突不破 30-40%的原因。

另一方面，如果聽說雙方有類似的生活環境，“按字面”進行翻譯並不妨礙聽的一方領會原話中的“言外之意、弦外之音”。

資訊檢索和文章分類

初級的工作可以只依靠“字串匹配”的手段進行。更高的要求則需要對關鍵字的內涵有一定的理解。例如，如果用戶要通過關鍵字“西部”和“大開發”檢索有關的文章，凡是含有“貴州”，“雲南”，“昆明”，“西安”……但是不包含“西部”的文章也不能放過。

直接執行

例如：按照電視劇腳本直接產生動畫片。在這種場合，“按字面翻譯”的手段就不能使用了，必須對腳本中的內容有真正的理解。

提取摘要

也許這種應用對“文化程度”的要求最高，因爲他不僅要求對文章的內容有所理解，而且能分辨出那些是無關緊要的廢話而加以剔除。

趨勢預測

基於語義，基於理解的自然語言處理 顯然是一項長期的戰略目標。這個目標一定會達到。如果把爲了達到這一目標所需要克服的困難比作一根蠟燭，那麼，這根蠟燭一定要“兩頭燒”才能燒得完。換句話說，在接近這一目標的過程中，會出現兩種潮流的匯合：

第一個方面是技術層面上的。電腦處理能力的提高，與自然語言處理有關的“基礎建設”的建立，都屬於這一類。這個方面是大家都認識到，天天都在討論的。

第二個方面是技術進步對語言本身發展的“反作用”。從歷史上看，書寫工具和書寫介質的變化對漢字形狀的影響是衆所周知的。那麼，電腦一歷史上前所未有的、如此大規模的應用也必然會在語言進化的方向上產生重大影響。（國家標準 GB2312 的頒佈和使用對老百姓爲後代取名字的影響，網蟲門聊天時發明的網路語言就是兩個鮮明的例子）。

“書面方言”的啓示

大家都知道，由於居住在不同地域的人群的口音的不同形成了地理意義上的“方言”。實際上，書面語言中也存在類似現象：嚴密的法律文書、簡潔扼要的商業信函、古色古香的尺牘大全、清新活潑的散文文體、Internet 出現以後所產生的網上聊天語言、乃至政府機關中使用

的公文“八股”，……。“八股”在這裏不是作爲貶義詞來使用的：一個剛剛畢業分配到機關工作的大學生所草擬出來的“學生腔的公文”一定會受到他的上司的堅決糾正。其內在的深層次的道理就是：同樣的意思只有通過同樣的、多次反復使用過的、爲人人所熟知的形式表達出來，才能讓文件的批閱者和執行者節省時間，不至於因爲出現了一些新鮮然而並沒有新的含義的字眼或者說法使他們產生誤解、白琢磨而耽誤工夫。

觸角無處不到的因特網上流動的文字資訊雖然是供人來閱讀的，但是由於這種文字信息量越來越大，今後它們將會越來越要依靠電腦來處理（）。因此，這些無處不在的電腦會在今後的社會中擔任自然語言的第二大（僅次於人的）用戶。可以預見，同時便於人和電腦理解和處理的**兩栖語言**極有可能會迅速發展起來，並且形成新的一種“書面方言”。

由於這種**兩栖語言**首先應該在因特網上得到應用，因此必須有一個統一的規範。包括：

- **句典** 這個語言不是由“規則”，而是由包含許多‘句式’或者‘模板’的“句典”來定義的
- 一套爲社會所公認的**標誌** 用來指明每一‘句式’或者‘模板’中可能存在的“語塊”的語義角色，語法性質
- 反映上述**標誌**系統的種類和組織關係的 XML DTD 文件。這一 DTD 文件屬於一種行業標準，以便於開發相應的 XML 解析器（parser）。這將有利於推動於中文資訊處理有關的應用開發。
- 最後，當然應該有軟體工具來幫助用戶書寫（改寫）出符合規範、並帶有共電腦處理的標誌。

可以認爲，前述北京資訊工程學院在承擔國家“九五”重點科技攻關專案“**中文資訊處理及產品開發**”中所開發的“**受限漢語寫作器**”已經沿這個方向走了第一步。但是其中的**例句庫**所收入的幾千個例句還要經過語言學家的檢驗，並且進一步擴充爲至少有幾萬種句式的**句典**。其中所使用的**標誌**系統也還要經過衆多專家的認真討論和修訂。**寫作器**的智慧程度也需要大幅度提高。

許嘉璐同志近年來多次呼籲: 要漢語言學界 和 電腦學界 兩支隊伍緊密地結合起來，開展面向中文資訊處理基礎研究和應用研究。他在 2000 年第 6 期《中國語文》上撰文說：

> 要消除中文資訊處理的瓶頸，**首要的關鍵**是要 漢語言學界 和 電腦學界 兩支隊伍緊密地結合起來，開展面向中文資訊處理基礎研究和應用研究。
>
> 如前所述，中文資訊處理需要的，並不是現在漢語學界已有知識的照搬：有的方面**需要根據電腦的“能力”去總結漢語的規律**，在一定程度上，還**需要研究者拋開傳統語言學的固有習慣和方**法；有的方面則需要填補上已有知識的不足。

自然語言是千百年以來爲人際交流的目的而發展起來的。現在，人們越來越多地“上網”，電腦越來越成爲這個語言社會中的一個重要的參與者。人們就會自覺地或者在不經意中**根據**

電腦的“能力”去總結漢語的規律，甚至改造人們的語言習慣。

一方面，技術向自然語言的深層衝擊，另一方面，自然語言向現代技術靠攏，這兩方面的發展趨勢會和起來，將會是解決“基於語義，基於理解的自然語言處理”這一戰略任務的過程中的一個重要特徵。